# Map of Research Funding: Comparison of Networking and Embedding Graphs

**CDKD 2018**, Beijing, China

**Ting Chen**

Institutes Of Science And Development, CAS

CONTENTS

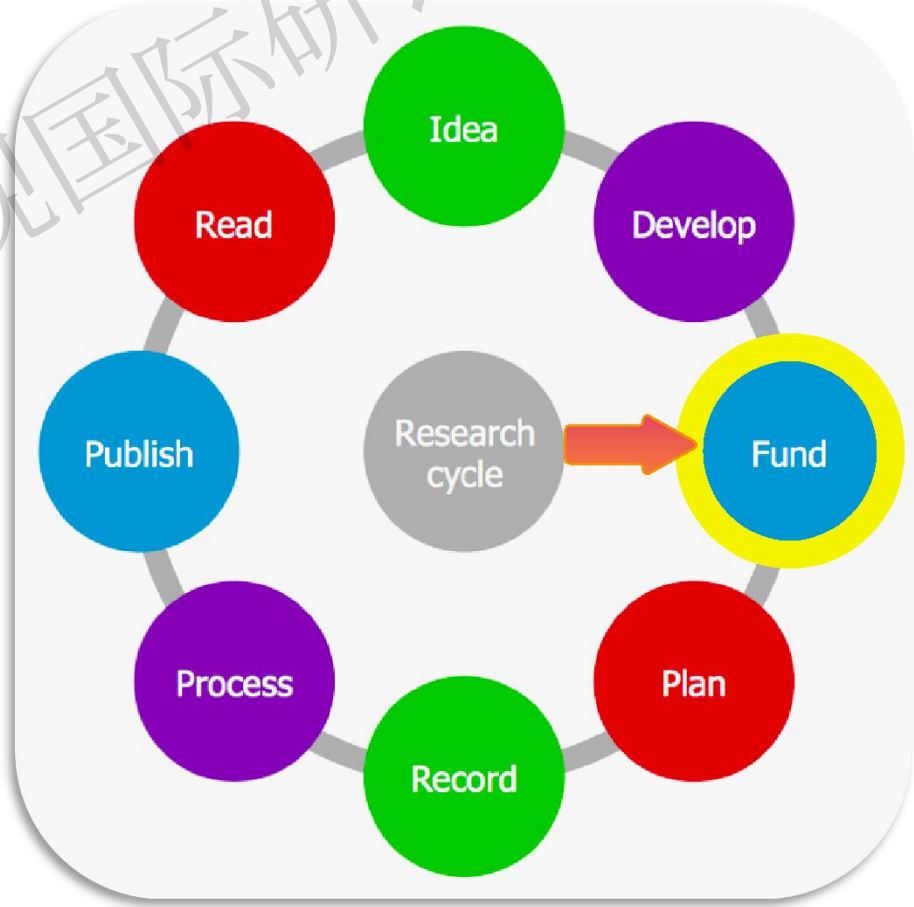
/01

# Introduction and Background

Why funding data ?

# *Why funding data ?*

- A typical publishing process takes 2 to 4 years from applying for funds to publishing a paper. Then another 1~2 years waiting for citations;
- Published papers focus on the research details (hard to understand), whereas funding applications are more about describing the ideas and direction (easy to read);
- Funded awards have also been peer-reviewed, but fund-related data never receives the same attention as papers and patents (Data and analysis methods are both lacking)



# *Main funding analyze methods*

**As an indicator**

The funded data is used as an indicator of innovation by counting different institutions and countries;

**Funded papers**

Analysis agency' funding layout by funded papers

**Clusters and Topics**

Using the textual features of application and clustering algorithm or the topic model divide awards into the research topics;

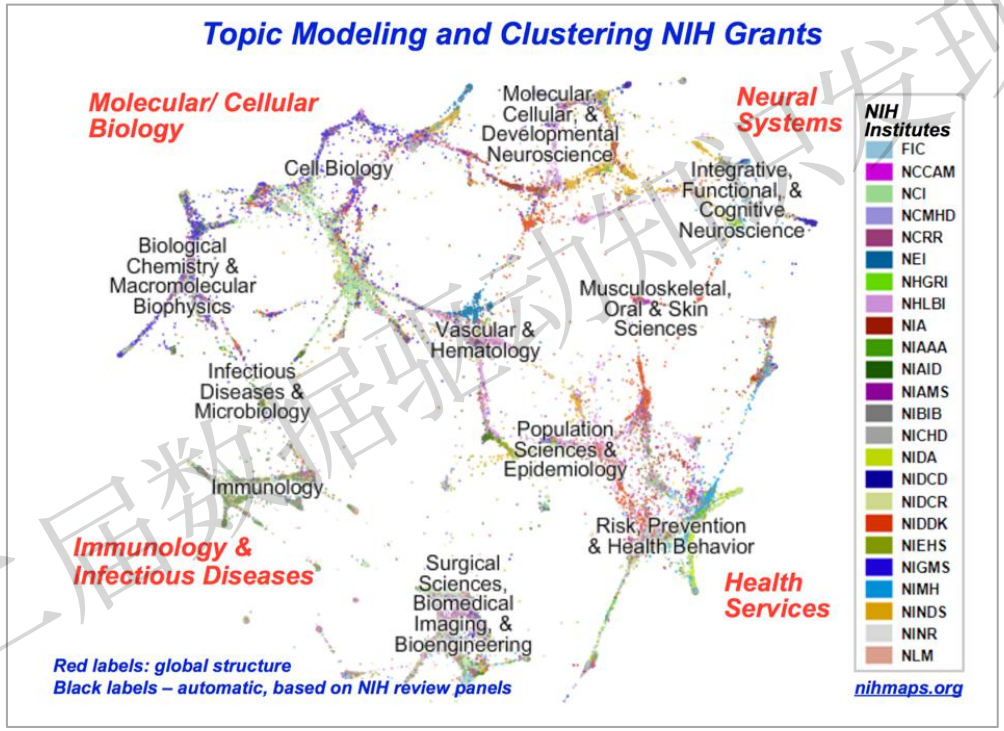
**There is often no visual display in existing funding analysis researches.**

# *Existing funding maps*

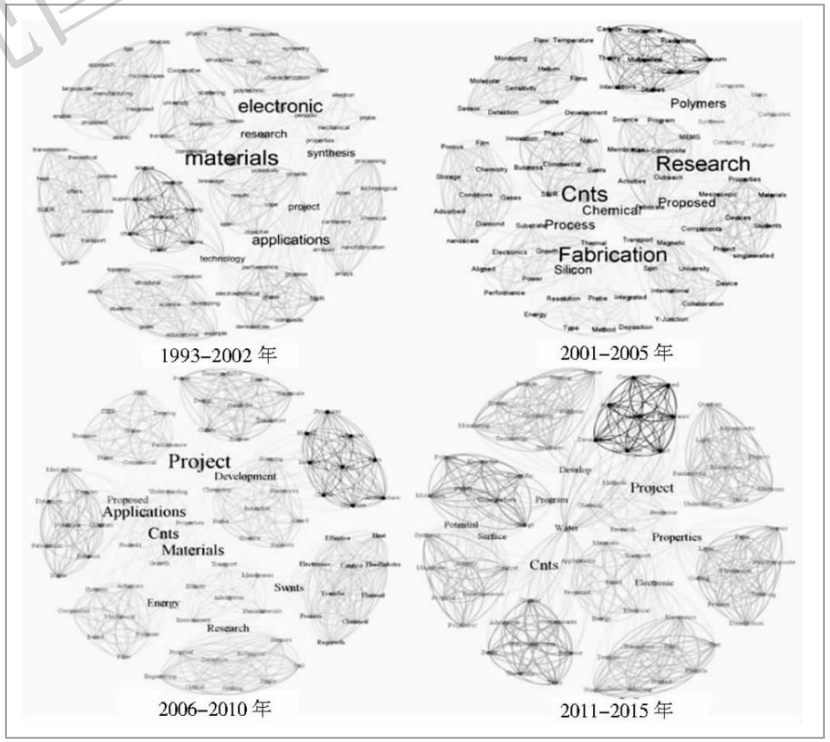
**Why map ?**

- Exploration
- Evidence
- Support

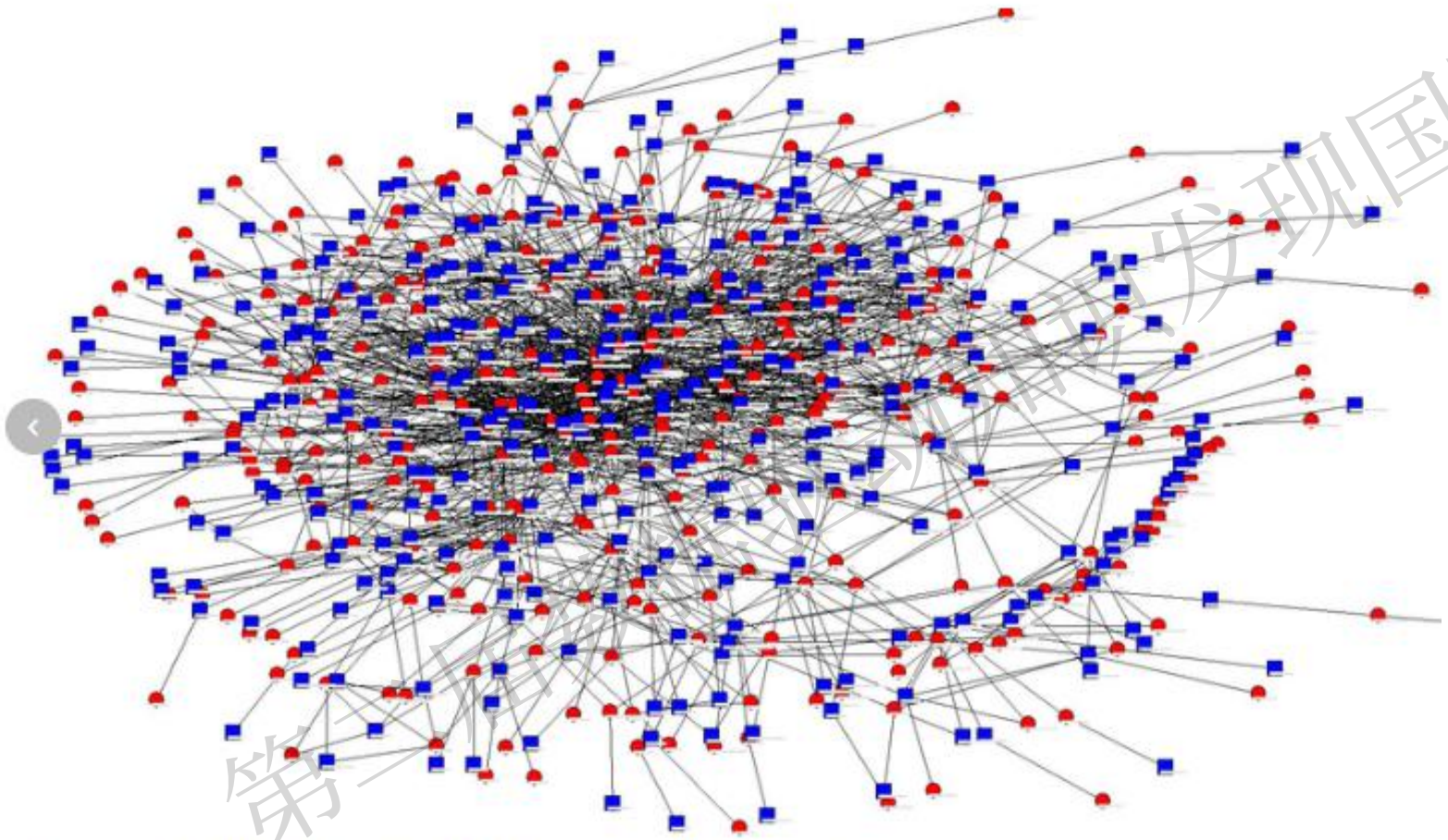
The NIH Visual Browser: An Interactive Visualization of Biomedical Research, IEEE conf, Information Visualization, 2009




The Method of Research Front Topic Detection Based on the Fund Project Data , LIS, 2017

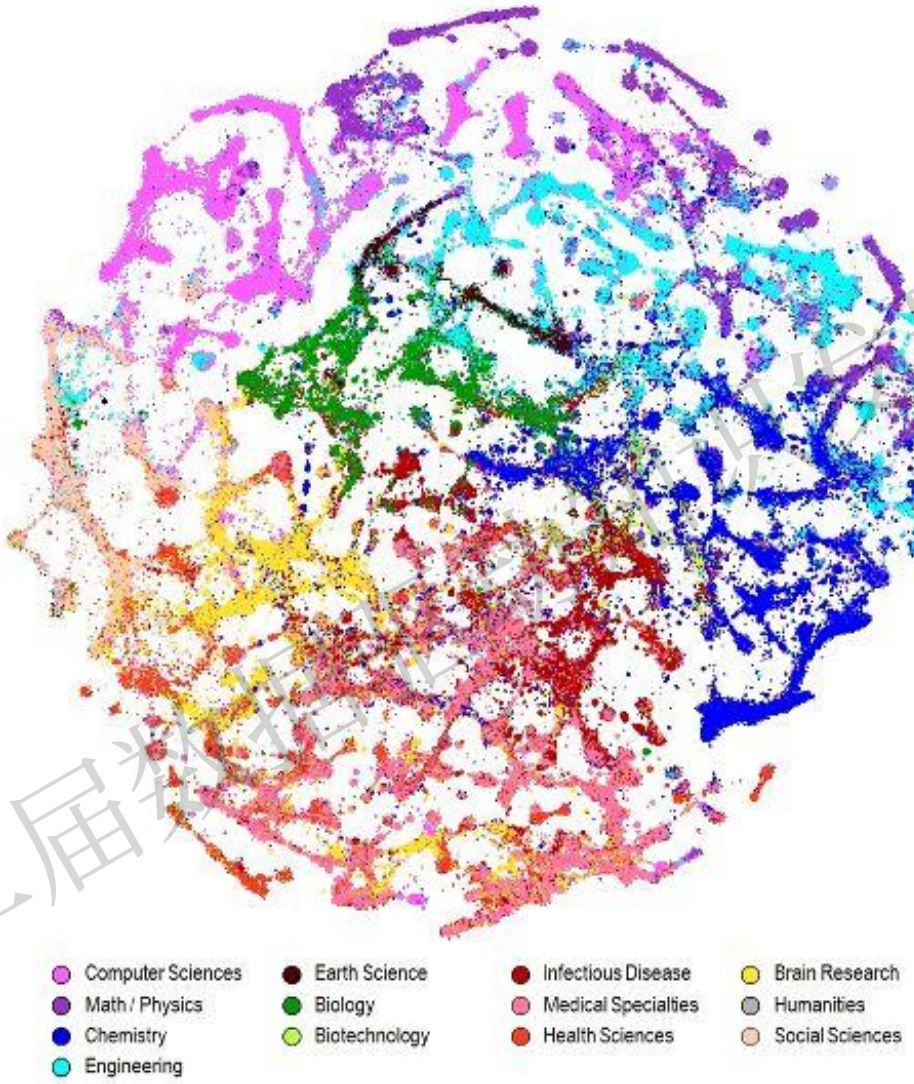


# What is a good visualization map?



The Co-Author Network Graph Of 511 Researchers Within Erdos1

# What is good visualization map?




by Boyack and Klavans (2013)

/02

# Create Labeled Test Dataset

We collected 4669 awards which were funded by NSF Information and Intelligent Systems department from 2008 to 2017

# *Labeled Test Dataset*

- Use k-means to divide 4669 awards into **70 small clusters**, smaller clusters, better **homogeneity**;
- Human-read each cluster, **combined similar clusters into one**, made sure the test set also had good **completeness**;
- **Total 21 topics** have been labeled, we will test our mapping methods by using the 21 topic labels. Topics include NLP, data retrieval, database, image recognition, voice recognition, motion monitoring, robotics, brain-computer interface, etc.

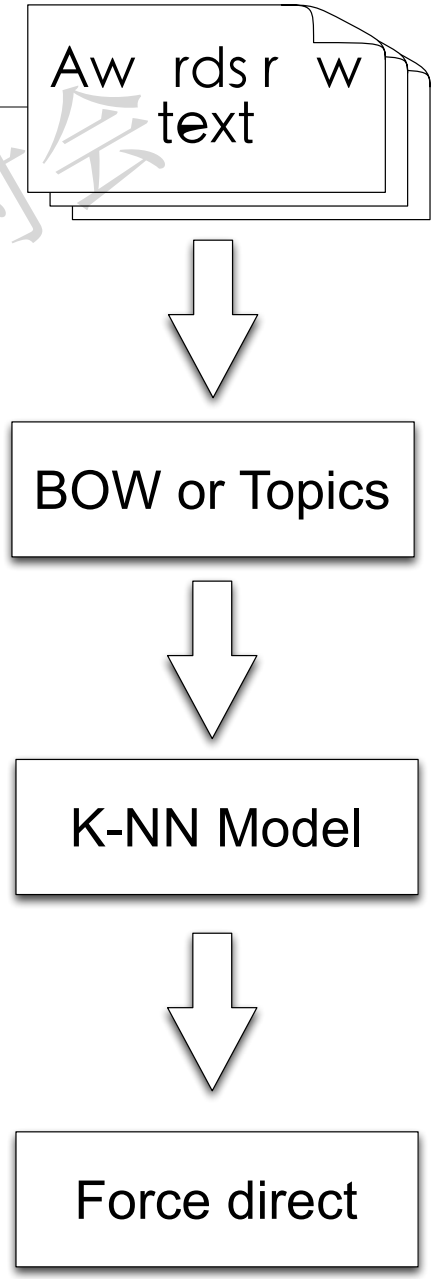
/03

# Networking or Embedding?

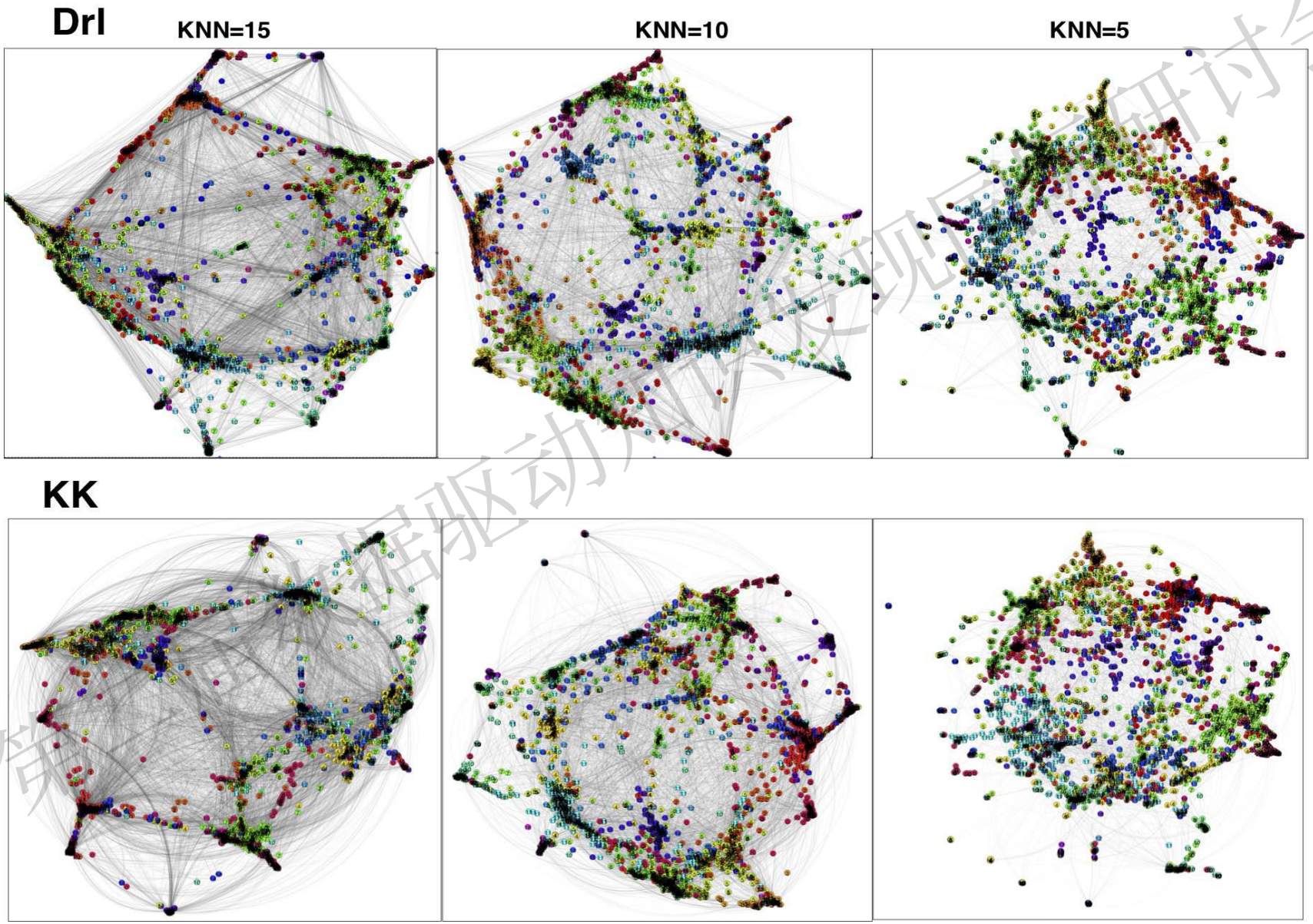
Compared two graph methods

# *First Try: Network Graph*

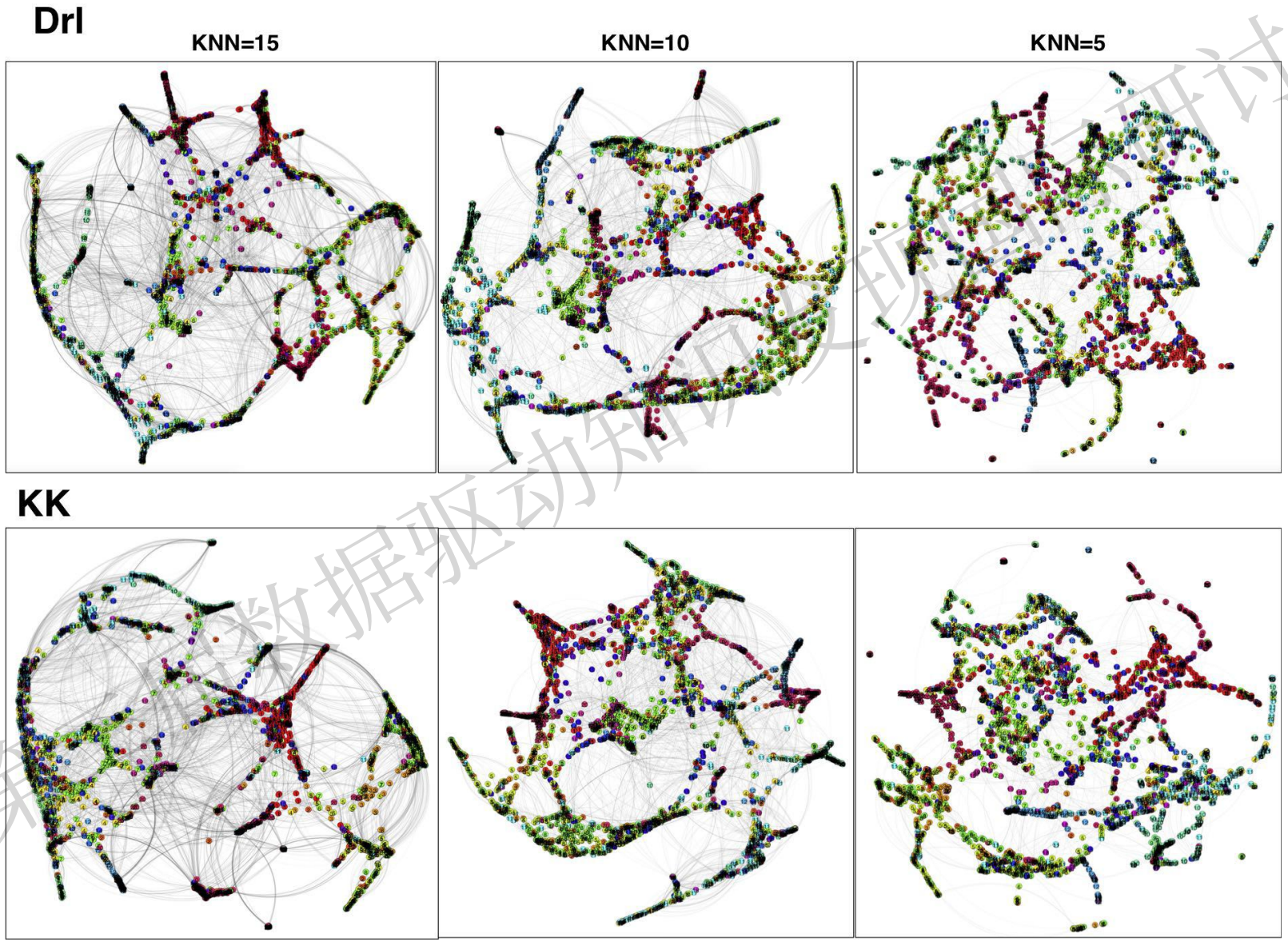
1. **Standard NLP:** Stop words, stemming and lemmatization, 2 or 3-gram phrase...;
2. **Feature extraction:** BOW tf-idf and topic model LDA;
3. **Create a network:** KNN model, K=5,10,15;
4. **Force direct layout:** Two most common medium-sized networks  layouts Drl (OpenOrd in gephi) and Kamada-Kawai (KK) were applied;



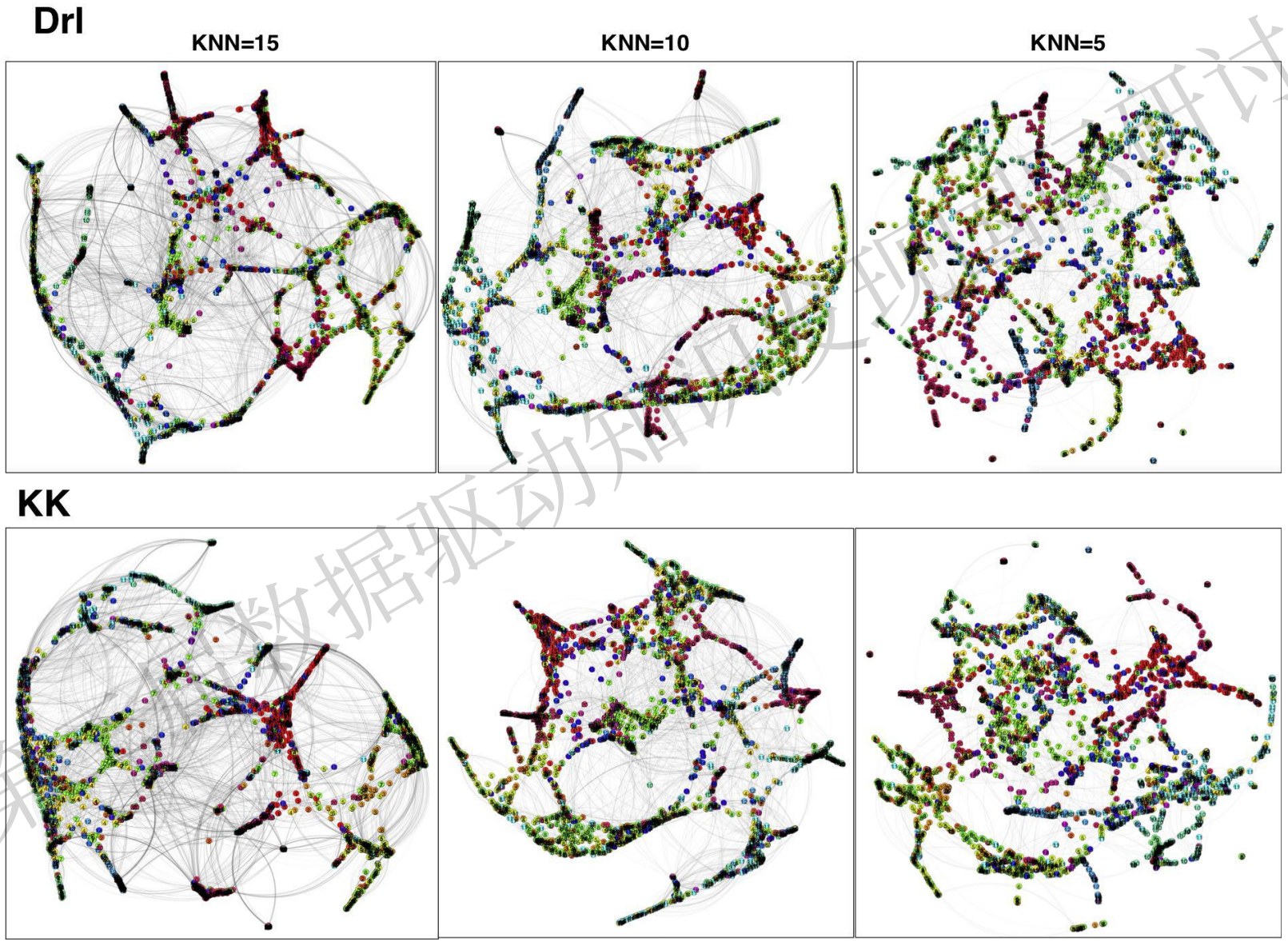
# *tf-idf similarity Graphs (7000 features)*



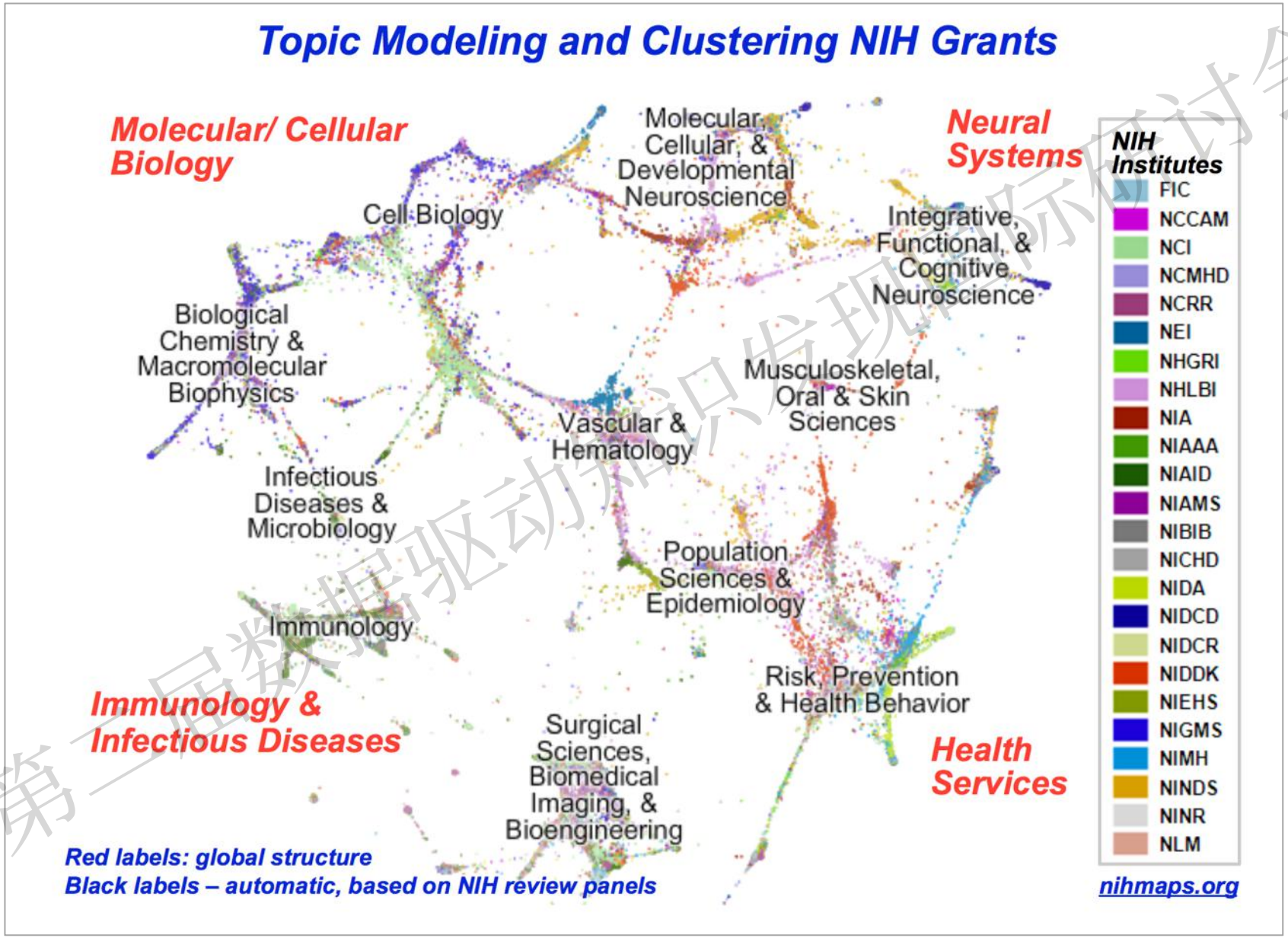
# *LDA similarity Graphs (20 LDA features)*



# *LDA similarity Graphs (20 LDA features)*



The NIH Visual Browser: An Interactive Visualization of Biomedical Research



**Best graph:**

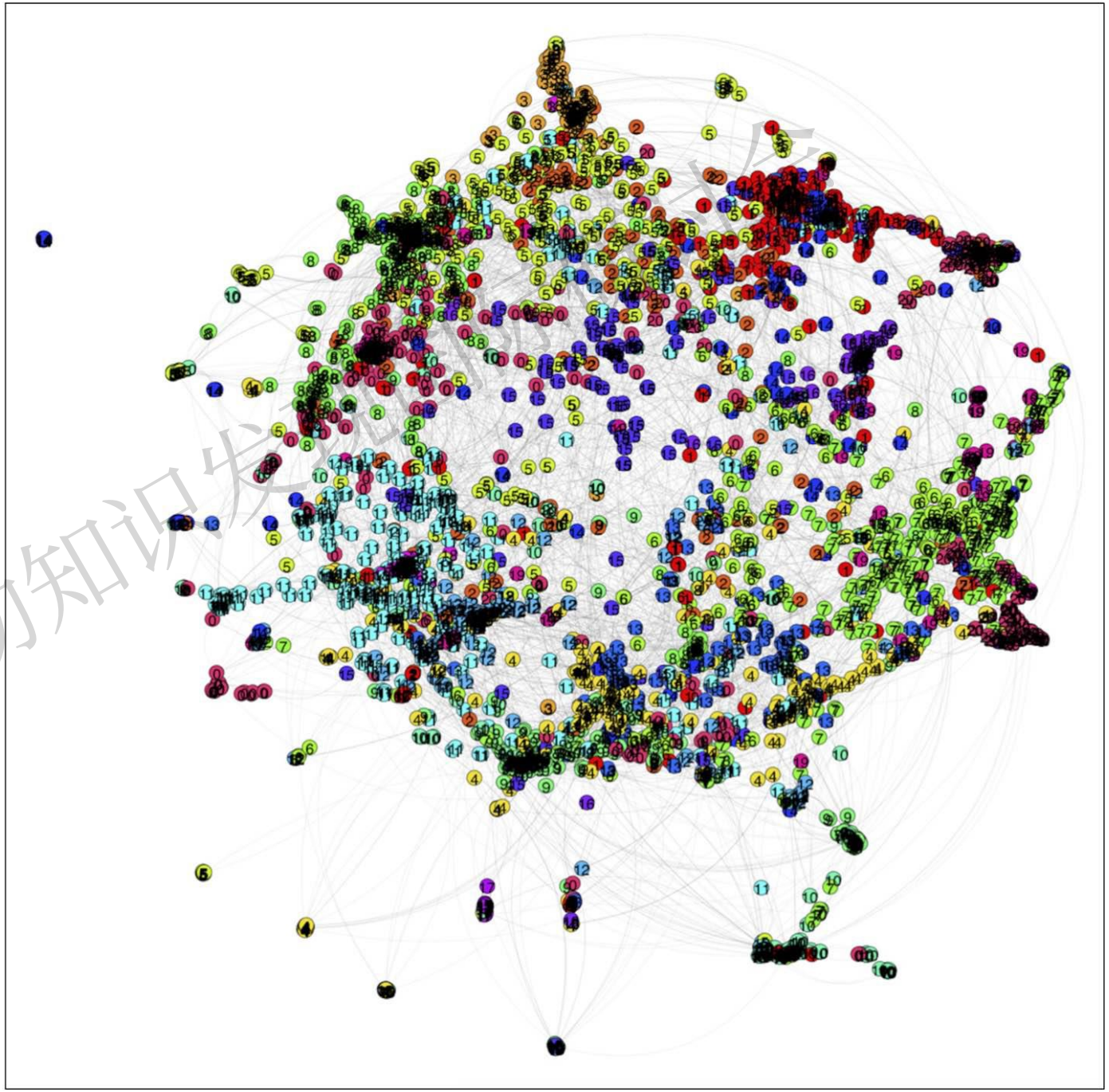
- BOW-Tf-idf  features
- K-Nearest neighbor K = 5
- Kamada-Kawai force direct layout
- Use it as a base map for this research

**Pros**

- Good global structure, some natural clusters appeared
- Degree, betweenness, centrality. etc
- Very fast

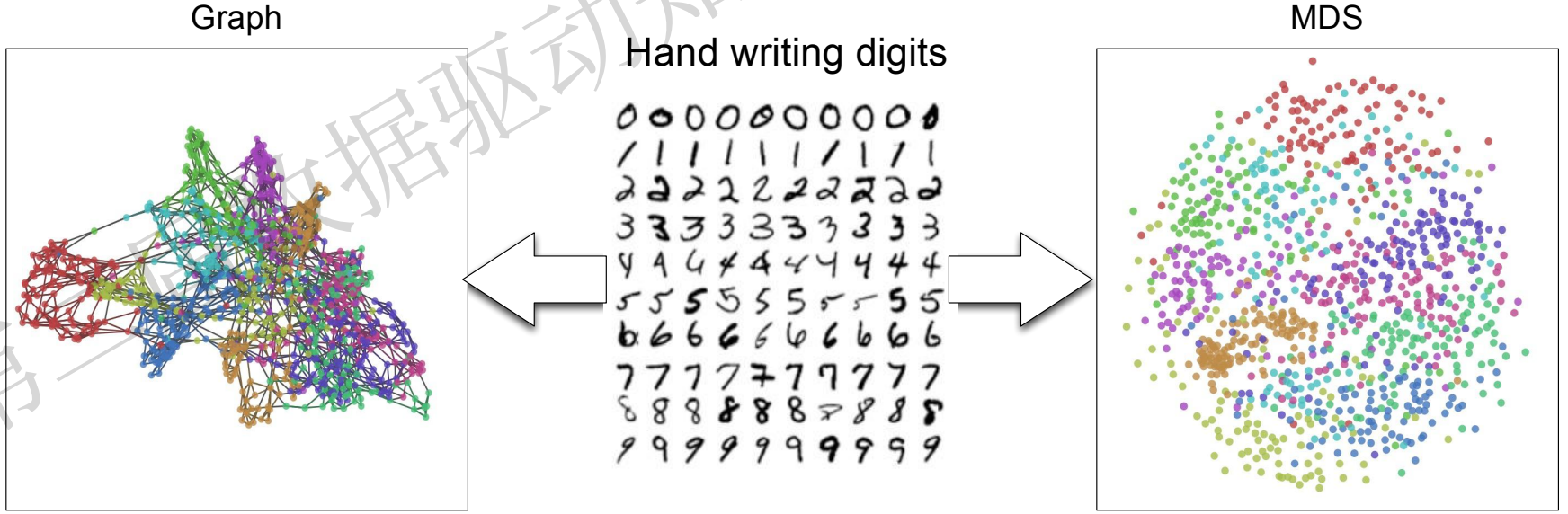
**Cons:**

- Poor topic-separability (the local detail)
- No real networks for funding data, we have to convert vector features into similarity  network (distance matrix)
- The choice of number of links is extremely critical



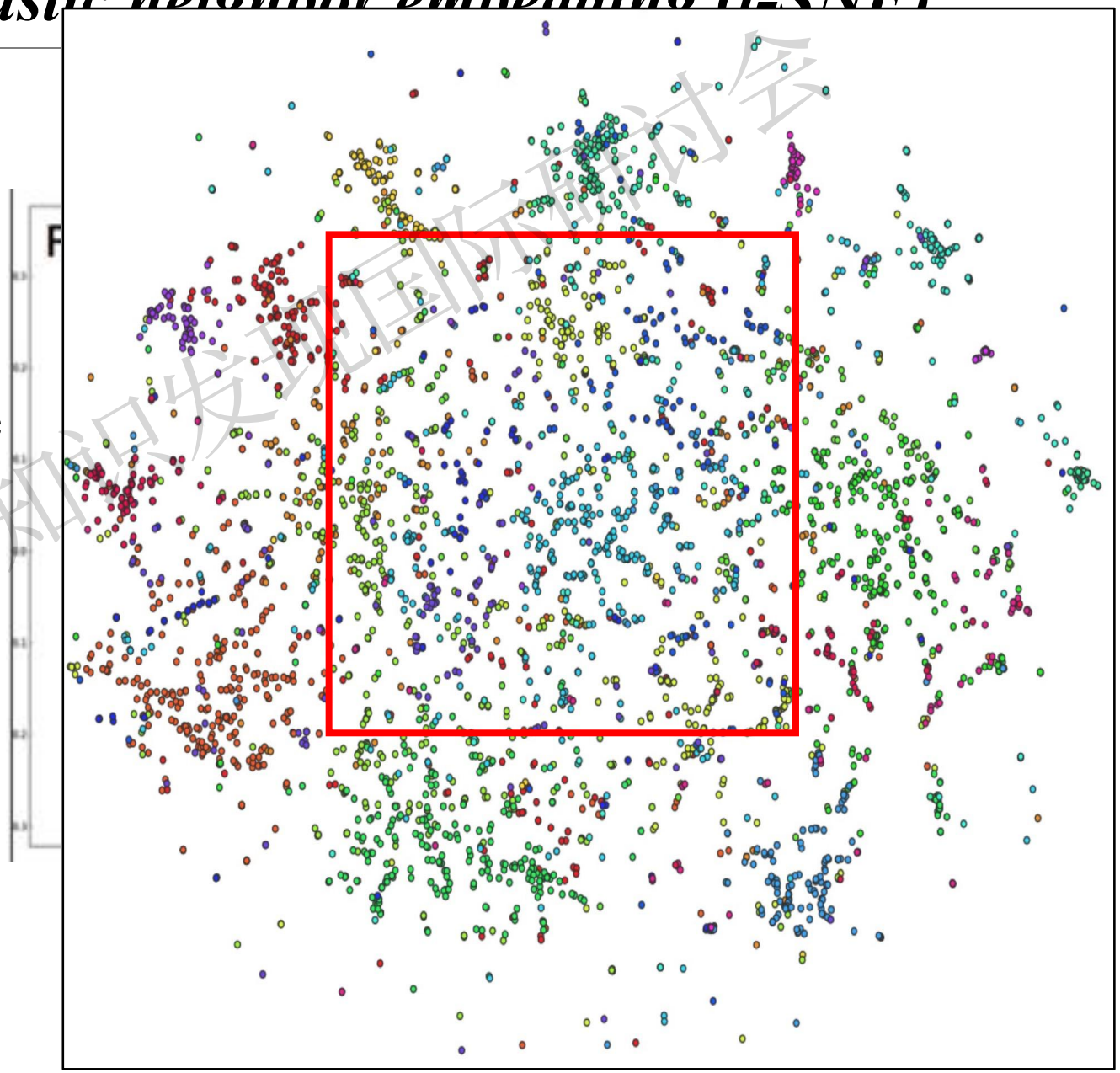
# *Second Try: Dimensionality Reduction (Embedding)*

- High-dimensional datasets can be very difficult to visualize. To aid visualization of the structure of a dataset, the dimension must be reduced in some way.
- Dimensionality Reduction methods were used for translating high-dimensional funding textual data into lower dimensional data;



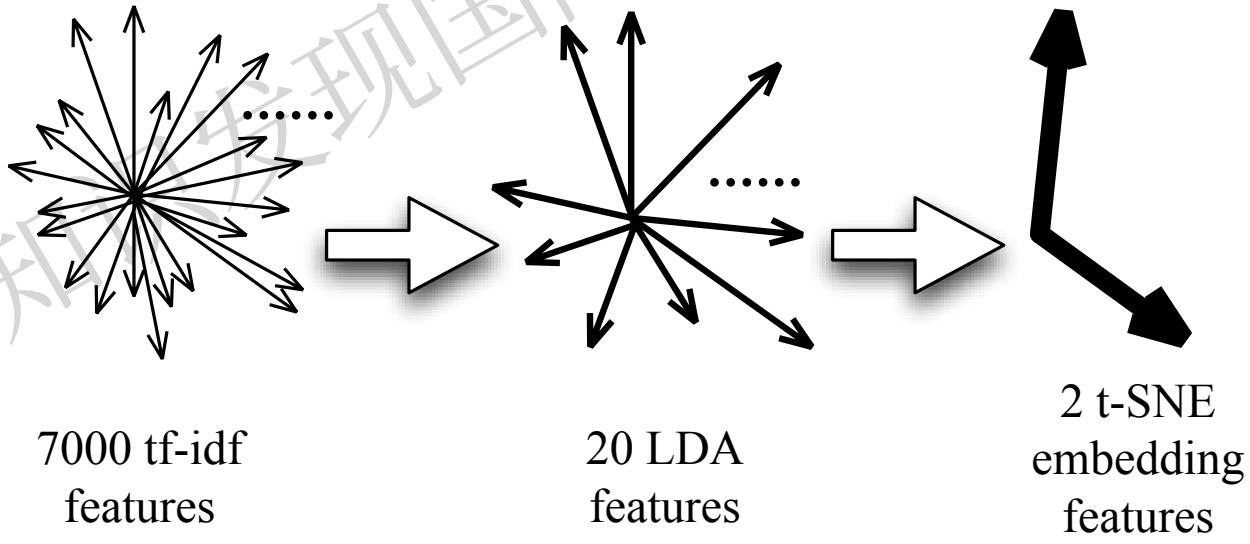
# *State of Art: t-distributed stochastic neighbor embedding (t-SNE)*

- Van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008)
- t-SNE tends to preserve local structure and at the same time preserving the global structure as much as possible
- Others try to preserve the global structure but missed a lot of local details



# *t-SNE + LDA topics*

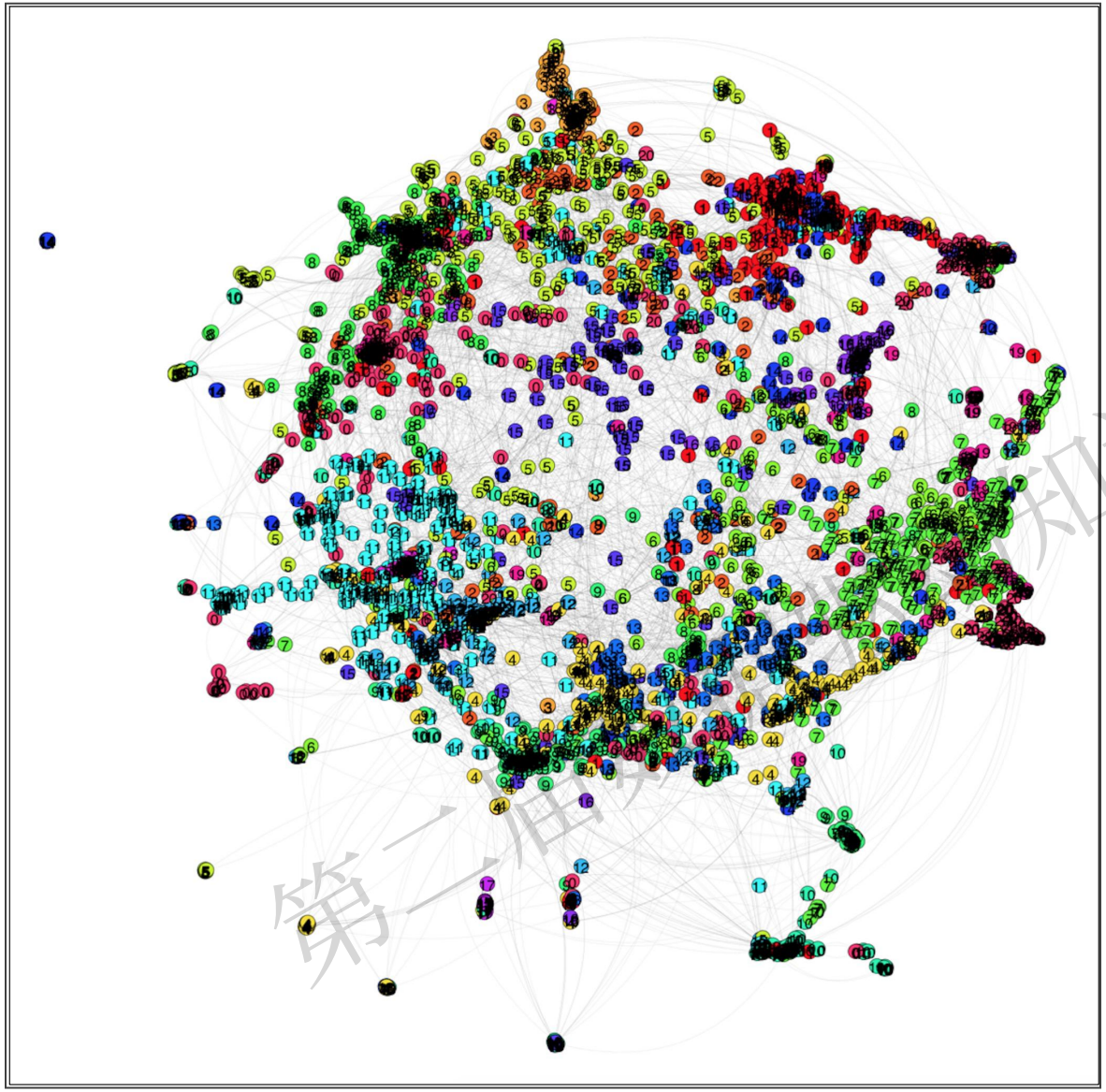
- 7000 tf-idf features are too high (Dimensional disaster);
- Add a topic layer between tf-idf and 2d space;
- Now, 20 topics features embedding to 2d space;
- well-separated clusters even in non-clustered data appeared on the map, even some sub-topics appeared in some larger cluster



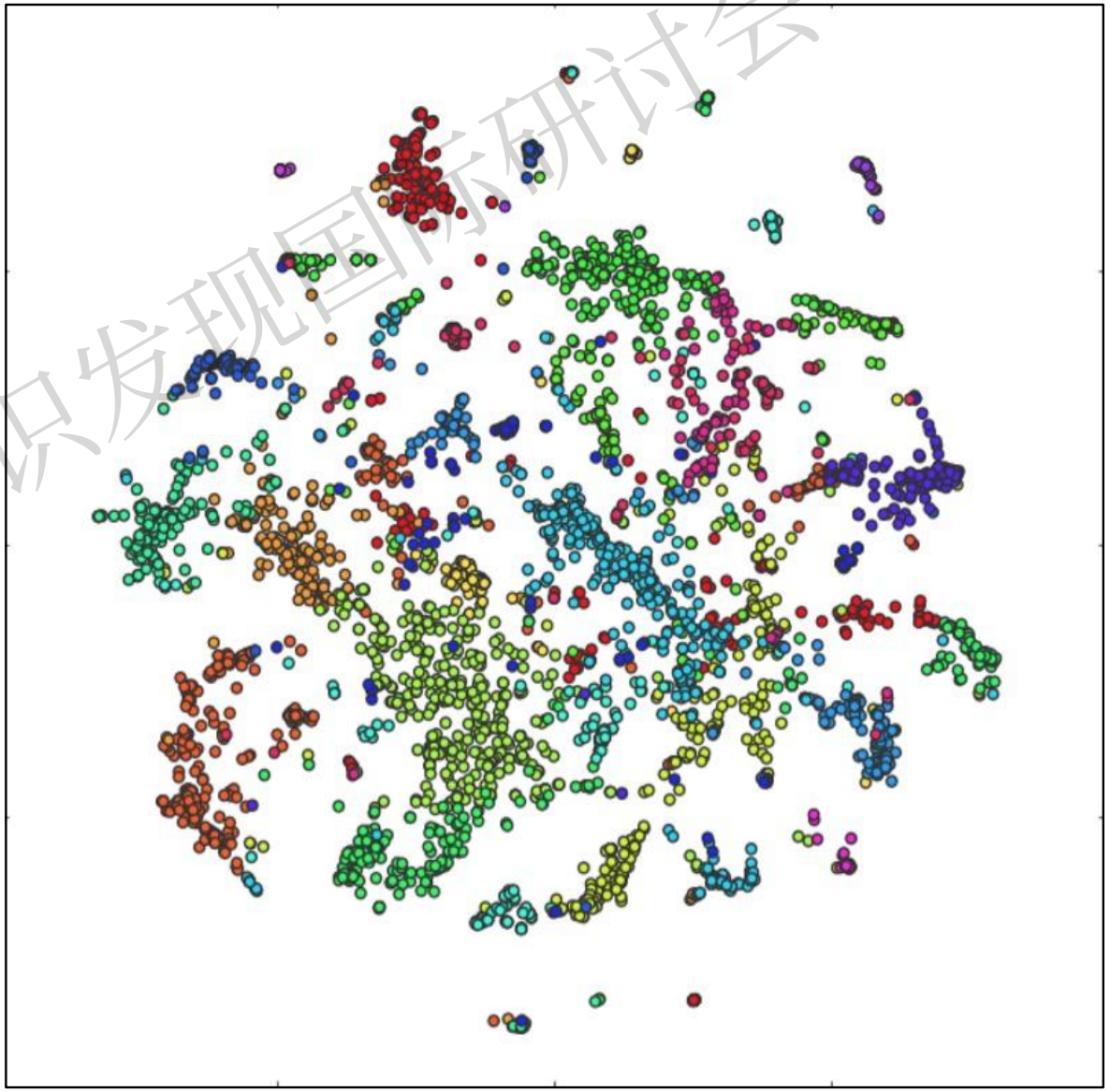


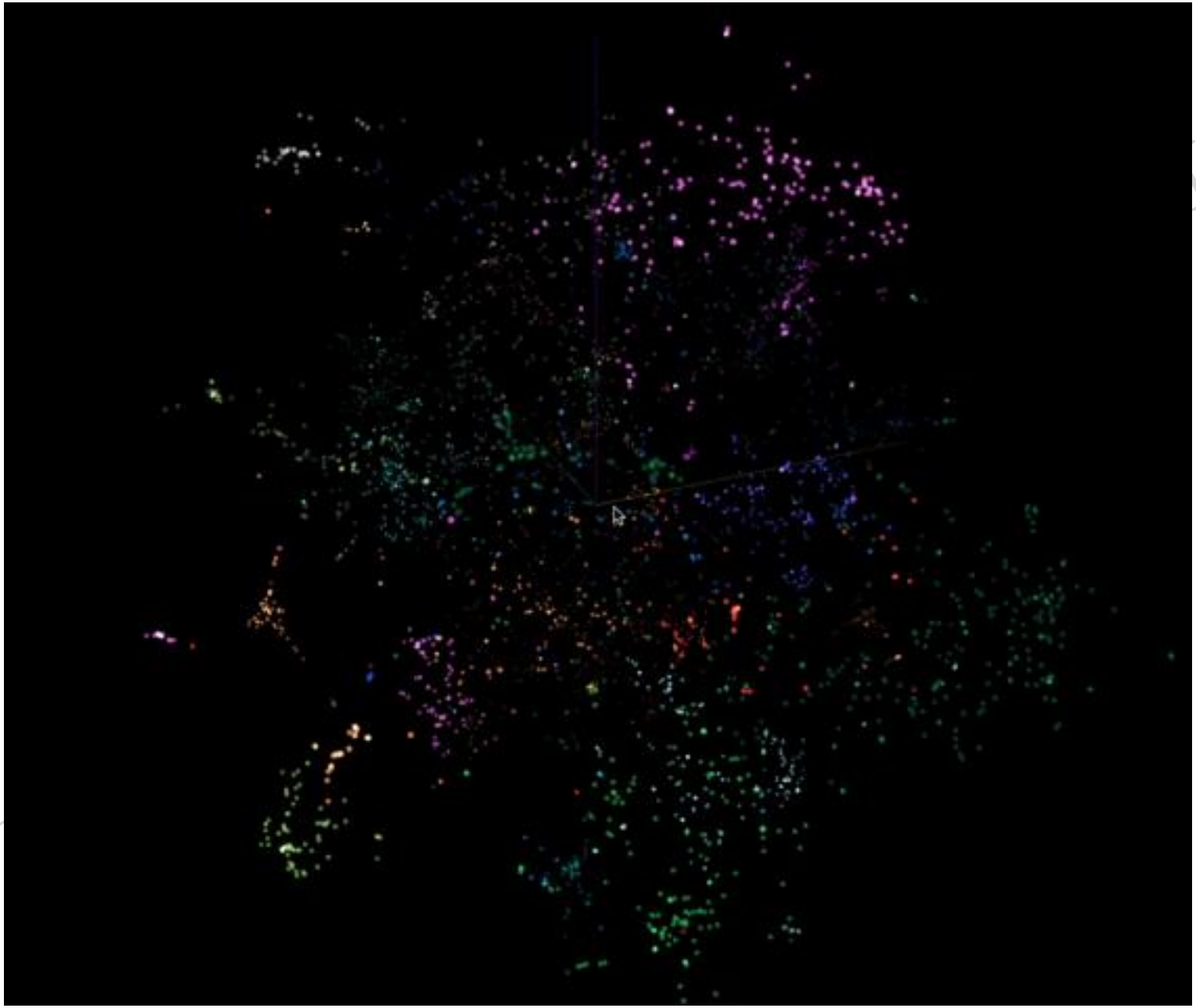

Ting Chen,Guopeng Li,Xiaomei Wang. Visualizing Appropriation of Research Funding with t-SNE Algorithm. Data Analysis and Knowledge Discovery, 2018.0251.

## Graph funding map (Base map)



## Embedding funding map





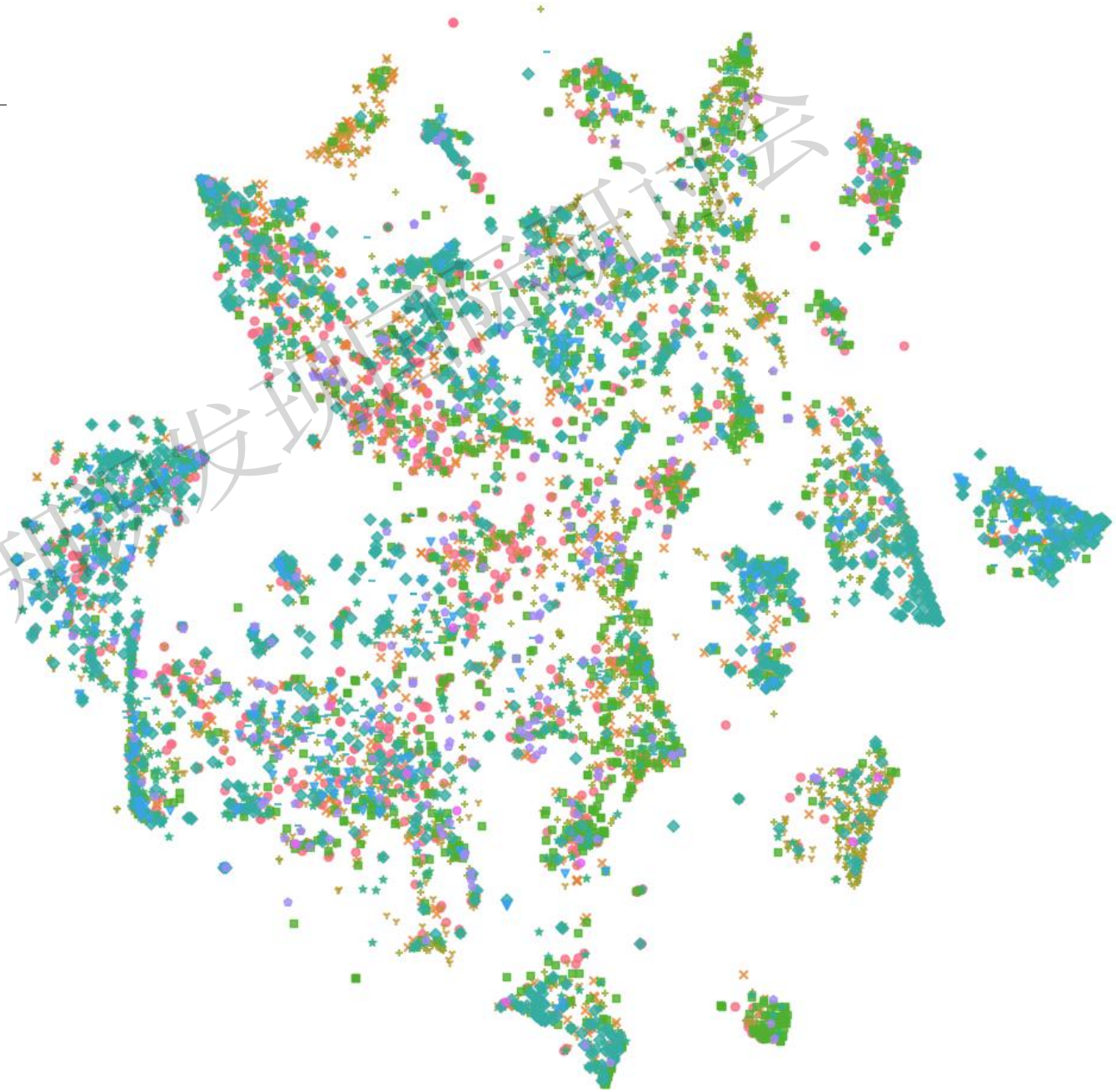
/04

# Applications of Funding Graph

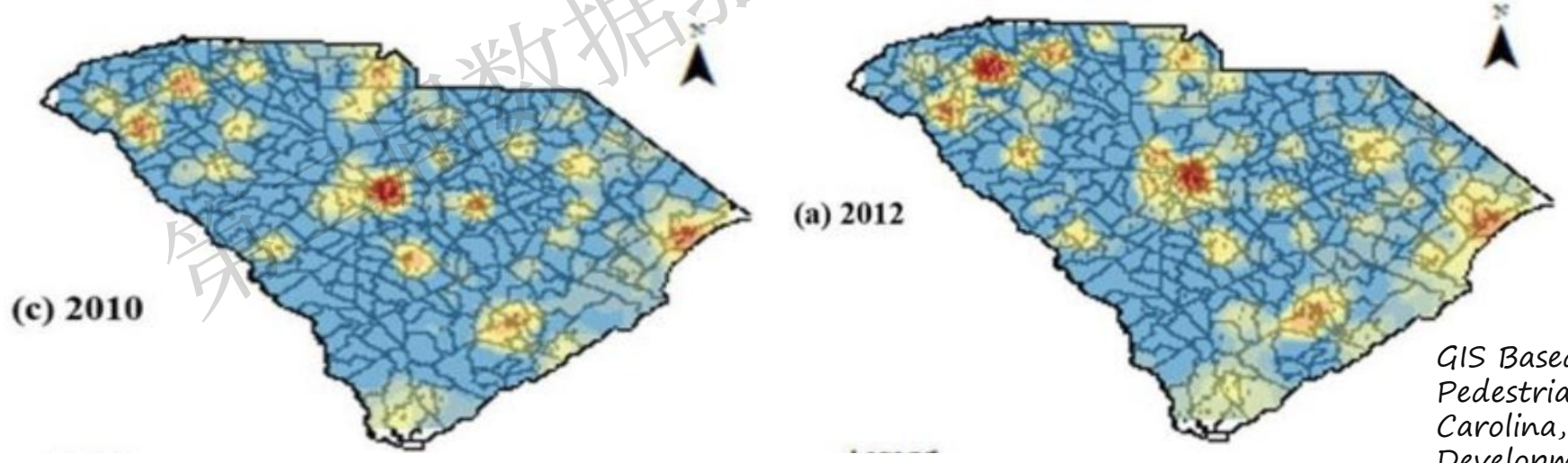
Hotspots detection; Novelty detection

# *NASA Funding hotspots*

- In the past 20 years, NASA SBIR has funded about 10,000 awards by ten centers;
- The funding map was created by using some method;
- We tried many clustering algorithms, hoping to divide awards into several clusters based on the map but the results were not good after experts interpretation.

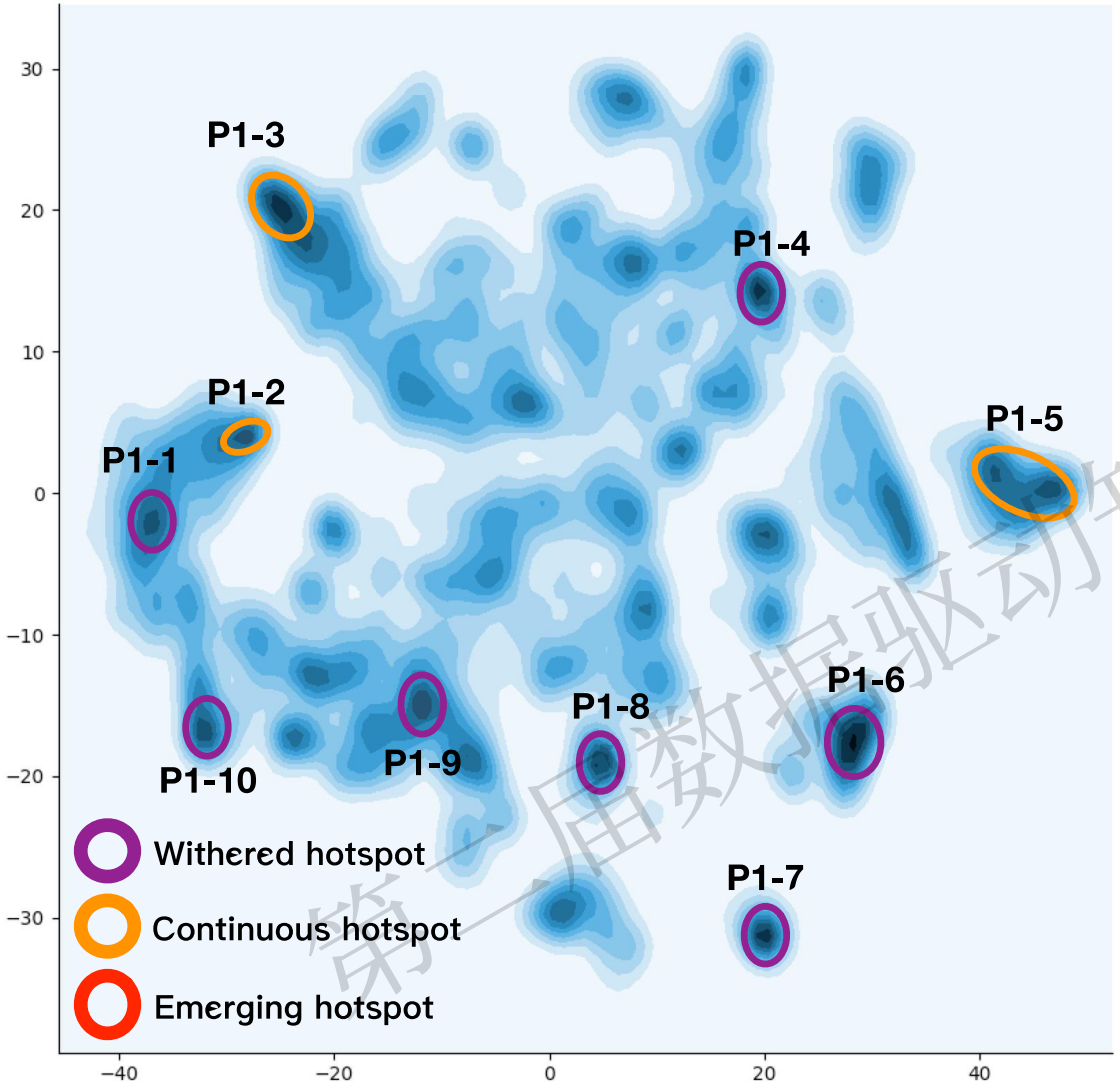


- The map provides good local structure, so we decided to skip the traditional clustering, instead of putting each item into a topic, we tried to find the region with the highest density in the map
- The high-density regions represent a large number of highly similar funded awards within a time period. We think they could represent the hotspots of research funding.
- The most commonly used method in GIS-the Kernel Density Estimation was used for finding hotspots on the map.

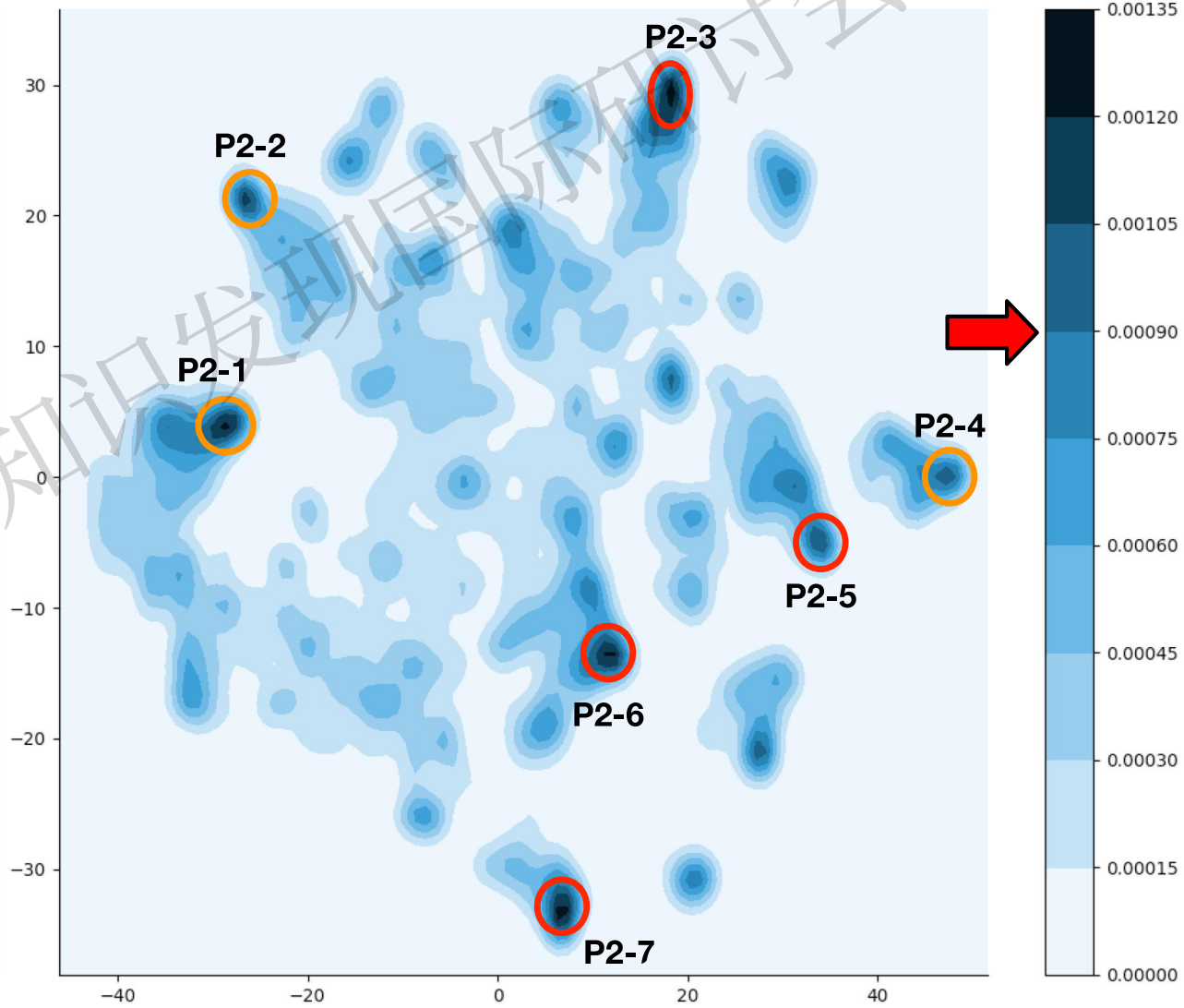



GIS Based Spatial Analysis of PedestrianCrashes: A Case Study of South Carolina, Conference on Transportation and Development 2018

## NASA SBIR funding map 2000 ~ 2008

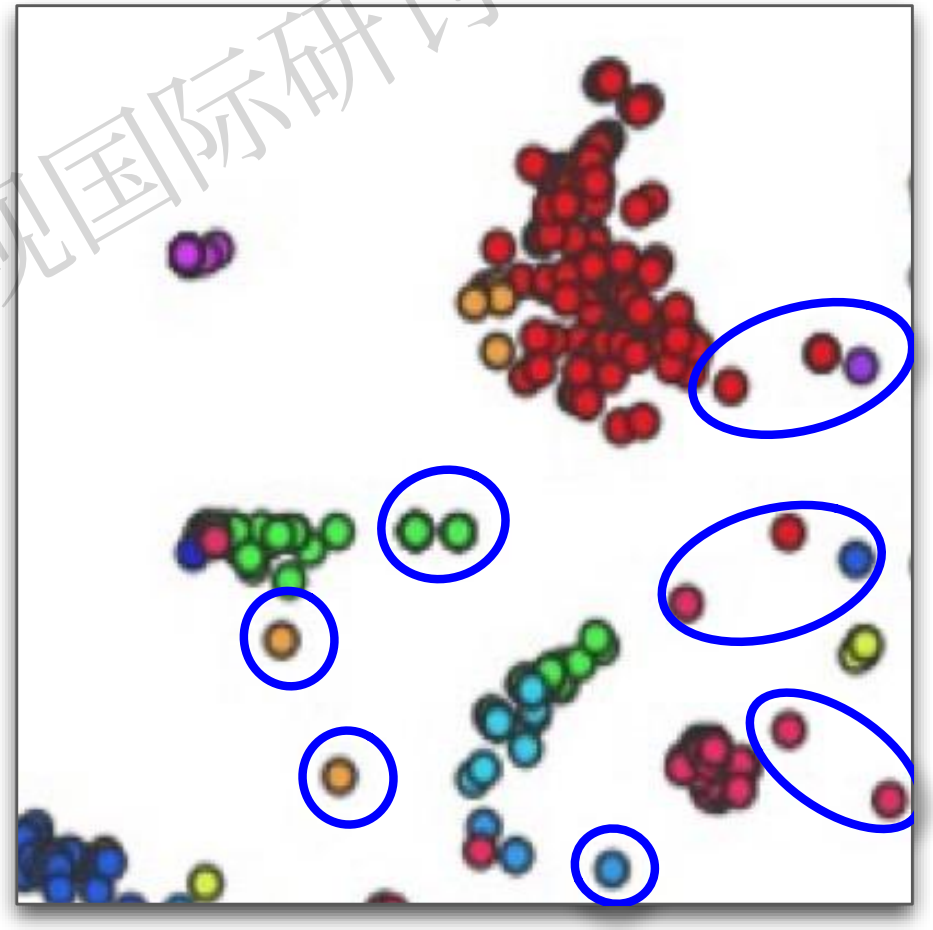



## NASA SBIR funding map 2009 ~ 2017



# *Novel application detection*

- Funding agencies are more likely to support novel research.
- Unlike the network graph, embedding won't lose the outliers (node without links), it will place the outlier at an appropriate position on the map even without any links.
- Some novelty/outlier detection methods can be used based on the funding map, such as one-class SVM, IsolationForest



/05

# Discussion

And the next step

# *Discussion*

- Both graph and embedding funding maps are good at revealing the global structure;
- The embedding map has the capability for retaining the local structure of the funding data, it seems to display natural clusters and sub-clusters very well;
- Text representing cannot be too high, better features will get a better map;
  - Tf-idf, BM25, LSA, LDA, NMF, Word2vec average/sum, doc2vec, Deep learning network
- The cost of t-SNE algorithm is O(n 2 ), not very fast and scalable;

# *Next step*

- Try to apply the funding map with multiple funding agencies' data, NSF/EURO Horizon 2020, maybe we will find some differences between counties and agencies;

- Test the method with other data sources, maybe patent or policy dataset;

# Thanks. Any question?


Ting Chen

chenting@casisd.cn